EDRN 2009:
Pitfalls to Biomarker Discovery:
Problems in Study Design

(adapted from talks at NCI BSA 2005, 2009)

David F. Ransohoff, MD

Division of Gastroenterology and Hepatology; Dept. of Medicine

Department of Epidemiology, School of Public Health

Director, Clinical Research Curriculum (K30)

Lineberger Comprehensive Cancer Center

University of North Carolina at Chapel Hill

“Validity”

Meaning of “validity” is broad (Lat: “strong”) and confusing; meaning must be clarified.

Nat Rev Cancer 2004; 4:309-14

Two critical threats to validity

1. **Chance**
   Does chance explain ‘discrimination’?

2. Bias
   Does bias explain ‘discrimination’?

Nat Rev Cancer 2005;5:142-9

The New England Journal of Medicine

VOLUME 347 DECEMBER 19, 2002 NUMBER 25

A GENE-EXPRESSION SIGNATURE AS A PREDICTOR OF SURVIVAL IN BREAST CANCER

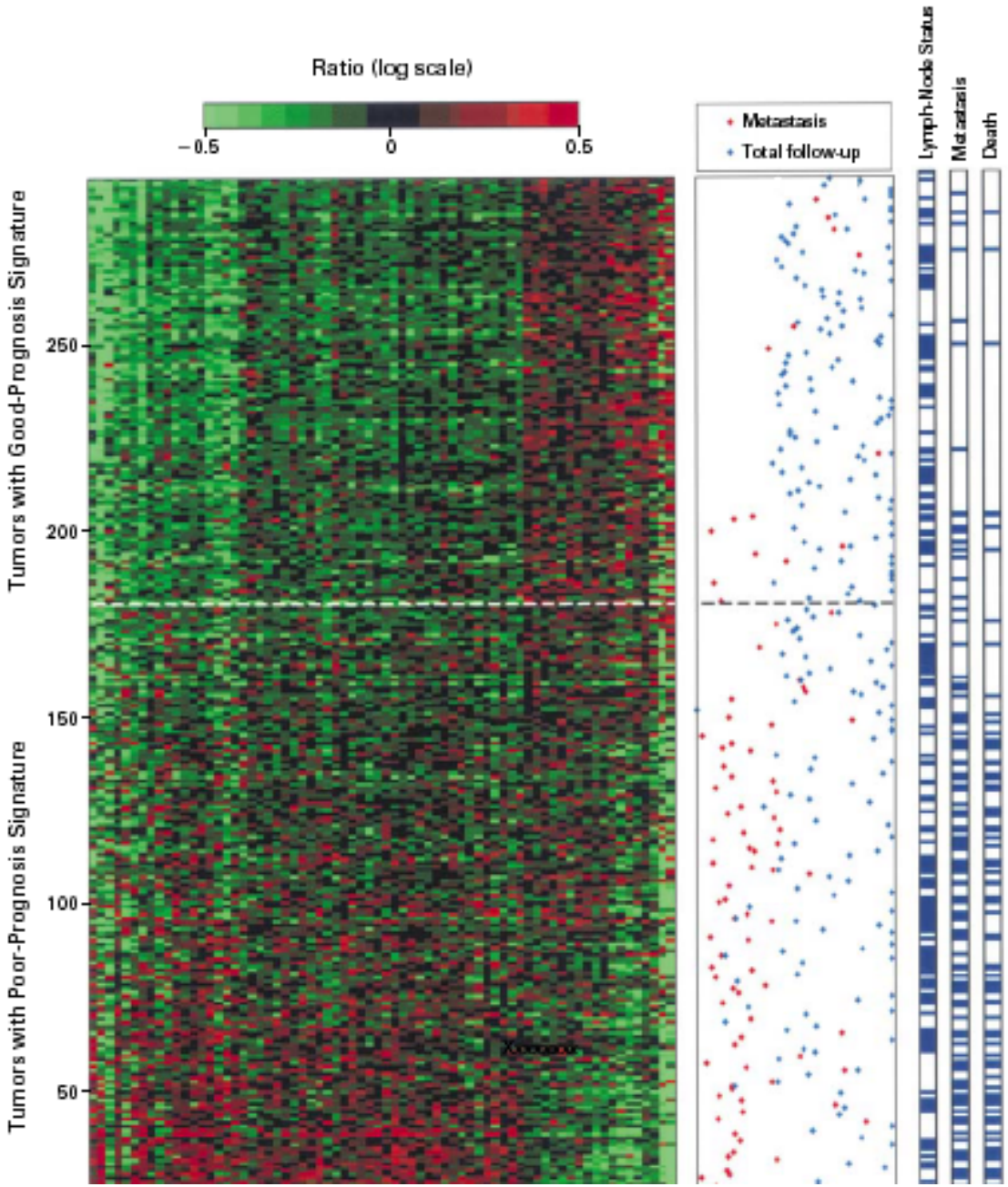
MARC J. VAN DE VIJVER, M.D., PH.D., YUDONG D. HE, PH.D., LAURA J. VAN 'T VEER, PH.D., HONGYUE DAI, PH.D., AUGUSTINUS A.M. HART, M.SC., DORIEN W. VOSKUIL, PH.D., GEORGE J. SCHREIBER, M.SC., JOHANNES L. PETERSE, M.D., CHRIS ROBERTS, PH.D., MATTHEW J. MARTON, PH.D., MARK PARRISH, DOUWE ATSMA, ANKE WITTEVEEN, ANNUSKA GLAS, PH.D., LEONIE DELAHAYE, TONY VAN DER VELDE, HARRY BARTELINK, M.D., PH.D., SJOERD RODENHUIS, M.D., PH.D., EMIEL T. RUTGERS, M.D., PH.D., STEPHEN H. FRIEND, M.D., PH.D., AND RENÉ BERNARDS, PH.D.

ABSTRACT

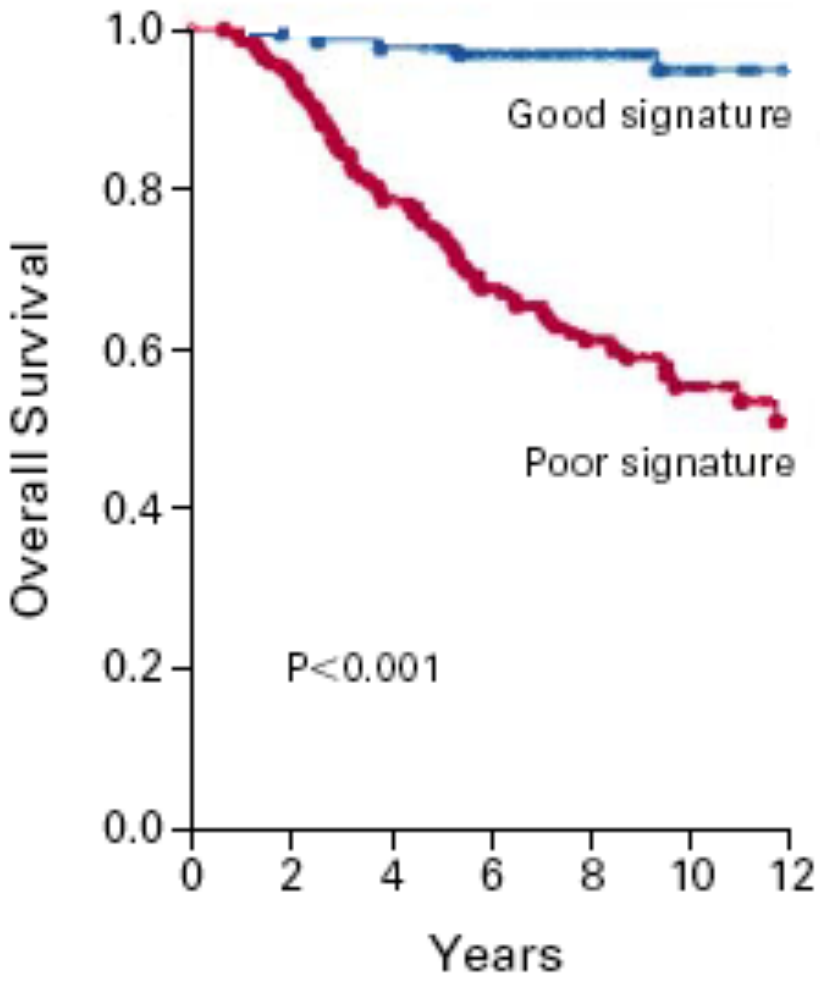
Background A more accurate means of prognostication in breast cancer will improve the selection of patients for adjuvant systemic therapy.

Methods Using microarray analysis to evaluate our previously established 70-gene prognosis profile, we classified a series of 295 consecutive patients with primary breast carcinomas as having a gene-expression signature associated with either a poor prognosis or a good prognosis. All patients had stage I or II breast cancer and were younger than 53 years old; 151 had lymph-node–negative disease, and 144 had lymph-node–positive disease. We evaluated the predictive power of the prognosis profile using univariable and multivariable statistical analyses.

ADJUVANT systemic therapy substantially improves disease-free and overall survival in both premenopausal and postmenopausal women up to the age of 70 years with lymph-node–negative or lymph-node–positive breast cancer.[1,2] It is generally agreed that patients with poor prognostic features benefit the most from adjuvant therapy.[3,4] The main prognostic factors in breast cancer are age, tumor size, status of axillary lymph nodes, histologic type of the tumor, pathological grade, and hormone-receptor status. A large number of other factors have been investigated for their potential to predict the outcome of disease, but in general, they have only limited predictive power.[5]


Ratio (log scale)
−0.5
0
0.5
Metastasis
Total follow-up
Lymph-Node Status
Metastasis
Death
Tumors with Good-Prognosis Signature
Tumors with Poor-Prognosis Signature
250
200
150
100
50

B All Patients



No. at Risk

Low risk	115	114	112	91	65	43	23
High risk	180	167	134	100	62	40	19

Strong discrimination led to interpretation as "*definitive*"

for clinical practice

"... gene-expression patterns of primary tumours are better than available clinicopathological methods for determining the prognosis of individual patients.[6,10,11]"

Ramaswamy and Perou, Lancet 2003;361:1576-7

for biological research

"... compelling evidence... genetic program of a cancer cell at diagnosis defines its biologic behavior many years later, refuting a competing hypothesis...."

Wooster and Weber, NEJM 2003;348:2339-47

Can chance explain results?

Definition: In multivariable predictive models, overfitting (a problem of 'chance') occurs when large N of predictor variables is fit to a small N of subjects. A model may 'fit' perfectly by chance, *even if no real relationship*.

(Simon, JNCI 2003)

Consequence: Results not reproducible in independent group.

Method to check for: Assess reproducibility in independent group.

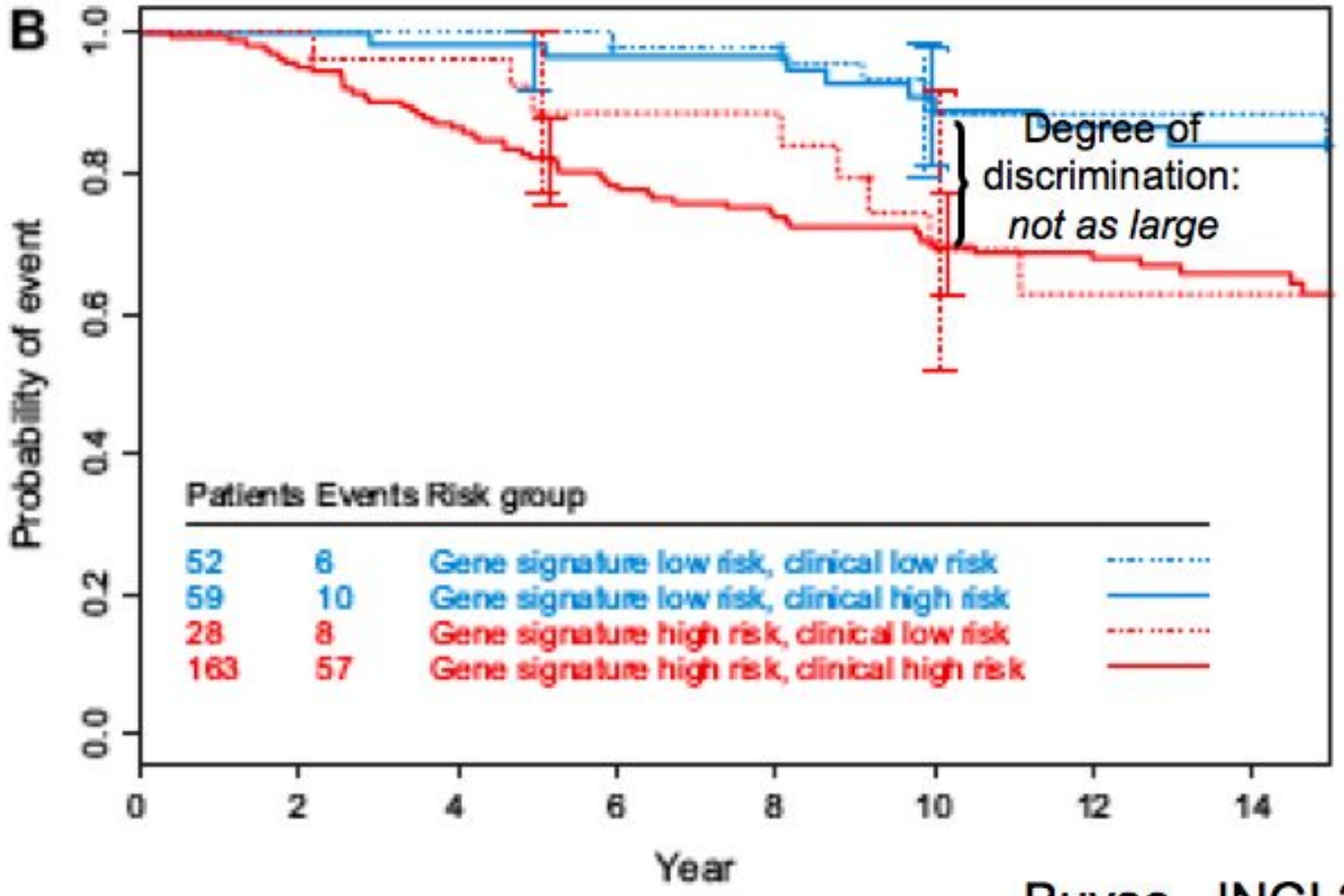
Can chance explain results?

to the editor:

"In research to validate a prognostic system, the inclusion of 61 patients from the… [training group in the validation group (N=295) means] the validation group is not independent.... [and] the degree of prognostic discrimination may have been inflated...."

(NEJM 2003;348:1716)

How much discrimination when different, independent subjects are assessed?



Buyse, JNCI 2006

If *less* discrimination, would interpretation be so strong?

for clinical practice

"... gene-expression patterns of primary tumours are better than available clinicopathological methods for determining the prognosis of individual patients.[6,10,11]"

Ramaswamy and Perou, Lancet 2003;361:1576-7

for biological research

"... compelling evidence... genetic program of a cancer cell at diagnosis defines its biologic behavior many years later, refuting a competing hypothesis...."

Wooster and Weber, NEJM 2003;348:2339-47

To check for overfitting, assess reproducibility in independent group

Nat Rev Cancer 2004;4:309.

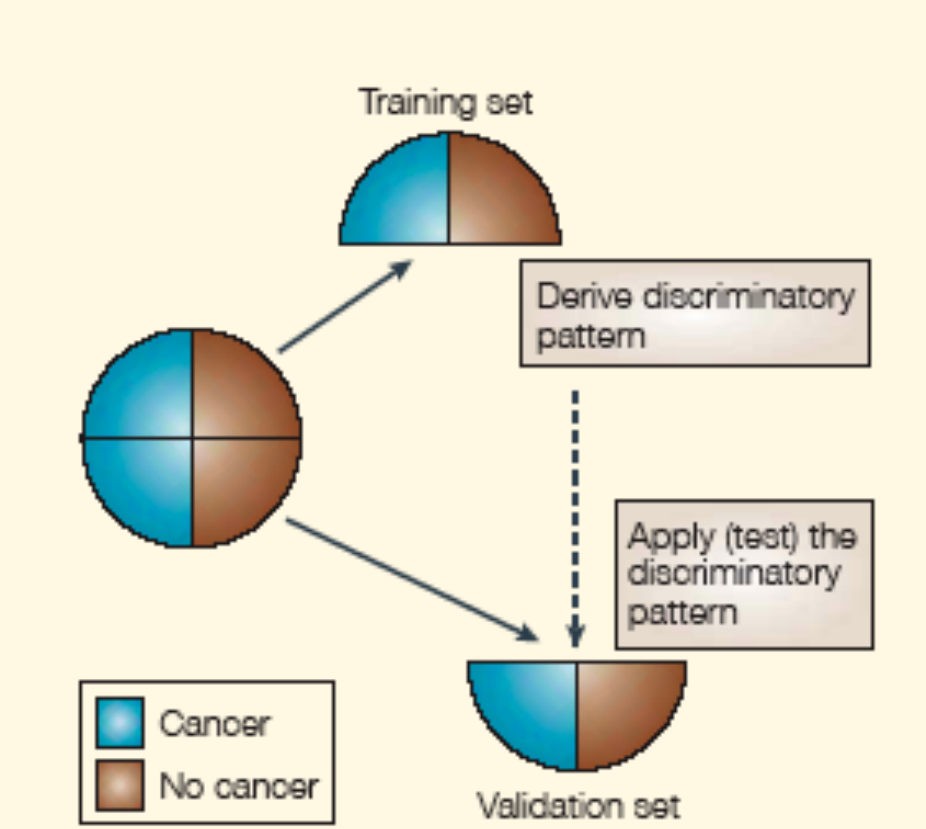


Figure 1 | **Method of dividing original sample to assess reproducibility and overfitting.**

Chance/overfitting *is* addressed in study of RNA expression

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer

Soonmyung Paik, M.D., Steven Shak, M.D., Gong Tang, Ph.D., Chungyeul Kim, M.D., Joffre Baker, Ph.D., Maureen Cronin, Ph.D., Frederick L. Baehner, M.D., Michael G. Walker, Ph.D., Drew Watson, Ph.D., Taesung Park, Ph.D., William Hiller, H.T., Edwin R. Fisher, M.D., D. Lawrence Wickerham, M.D., John Bryant, Ph.D., and Norman Wolmark, M.D.

N Engl J Med 2004;351:2817-26.

Chance/overfitting is addressed in study of RNA expression

... because Methods showed 'independent validation':

"The prospectively defined assay methods and end points were finalized in a protocol signed on August 27, 2003. RT-PCR analysis was initiated on September 5, 2003, and... data were transferred... for analysis on September 29, 2003."

N Engl J Med 2004;351:2817-26.

Chance as a threat to validity

Nat Rev Cancer 2004;4:309-14

OPINION

Rules of evidence for cancer molecular-marker discovery and validation

David F. Ransohoff

According to some claims, molecular markers are set to revolutionize the process of evaluating prognosis and diagnosis for cancer. Research about cancer markers has, however, been characterized by inflated expectations, followed by disappointment when original results can not be reproduced. Even now, disappointment might be expected, in part because rules of evidence to assess the validity of studies about diagnosis and prognosis are both underdeveloped and not routinely applied. What challenges are involved in assessing studies and how might problems be avoided so as to realize the full potential of this emerging technology?

described briefly, should be considered in similar depth.

Molecular markers

Reasons for optimism. Molecular markers hold great promise for refining our ability to establish early diagnosis and prognosis, and to predict response to therapy. Optimism about molecular markers is based on exciting new knowledge and new technology. Knowledge about the molecular biology of cancer allows the identification of candidate target markers, such as the mutations that occur during the evolution of colon tissue from normal to adenoma to invasive cancer[7,8]. Powerful technologies including POLYMERASE CHAIN REACTION, SERIAL ANALYSIS OF GENE EXPRESSION, SINGLE-NUCLEOTIDE-POLYMORPHISM analysis and MICROARRAYS can

temptation to be casual about adhering to rules of evidence could result in claims that are not reproducible and lead to disappointment.

Reasons for caution. Although molecular markers will undoubtedly provide advances in diagnosis and prognosis, the degree of success claimed at present is extraordinary. Will we look back in 10 years and find that initial results were not reproducible? In an example from a generation ago, carcinoembryonic antigen (CEA) was purported to be nearly 100% sensitive and specific for colorectal cancer screening in initial research[11], whereas subsequent research had very different results. History might not necessarily repeat itself, but it indicates caution before making claims of success[12]. The non-reproducibility of the CEA results was due, in large part, to the fact that individuals who were initially studied had extensive cancer, whereas individuals who were later studied had less extensive asymptomatic cancer in which CEA might not have been increased[13,14]. The fact that test results vary with the 'spectrum' of disease might seem obvious now, but there was little understanding in that era of the concept of spectrum and of the biases that affect research about diagnostic tests. Development of the methods and

Two critical threats to validity

1. Chance
   Does chance explain ‘discrimination’?

2. **Bias**
   Does bias explain ‘discrimination’?

Nat Rev Cancer 2005;5:142-9

Experimental design and biospecimens

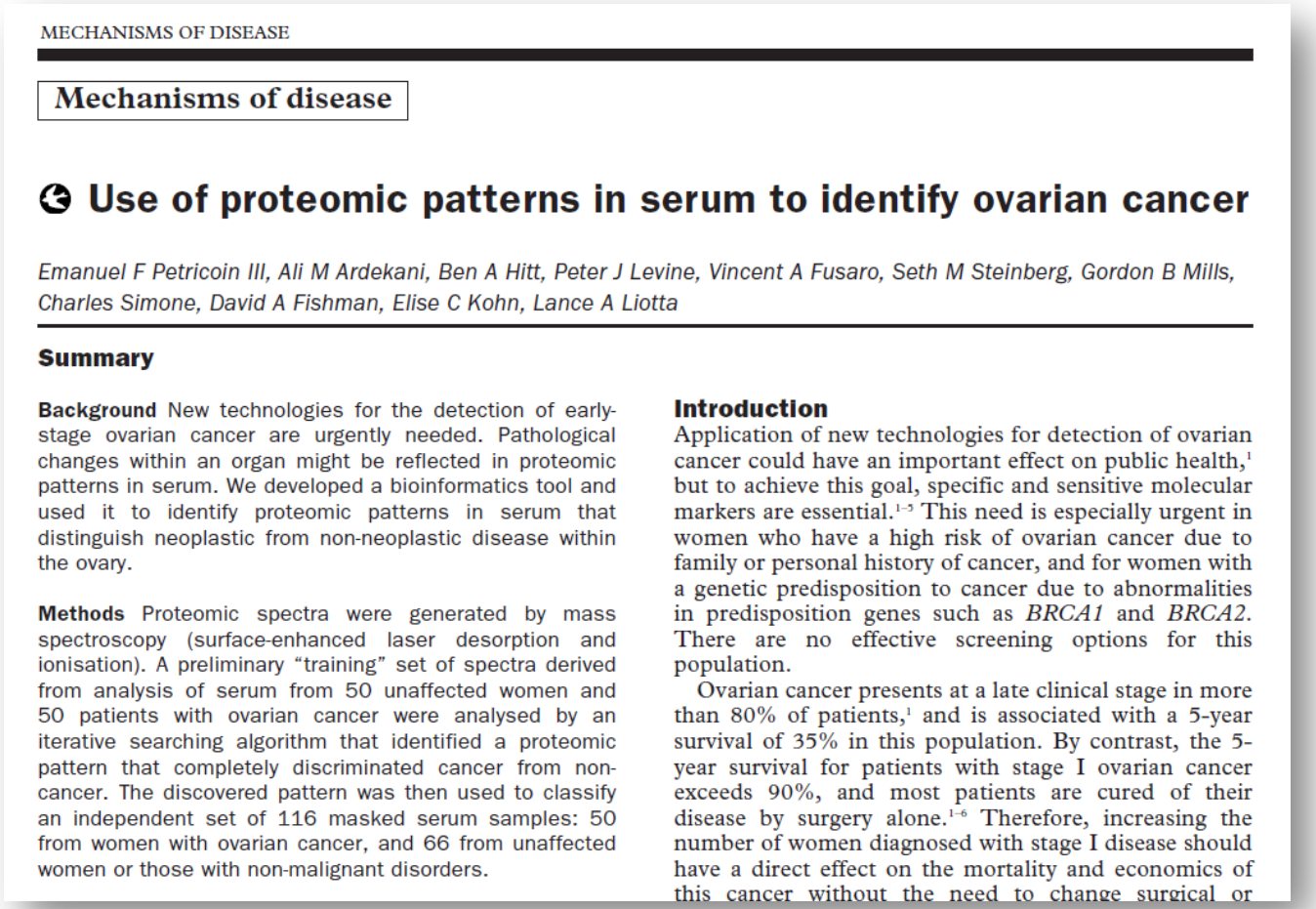
Problem

- In biomarker research, rate-limiting step is faulty study design, when bias (systematic difference between compared groups) makes results wrong and misleading.

Approach

- (to be described)

Problem: Bias – Example 1



MECHANISMS OF DISEASE

Mechanisms of disease

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

Summary

Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Methods Proteomic spectra were generated by mass spectroscopy (surface-enhanced laser desorption and ionisation). A preliminary "training" set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analysed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders.

Introduction

Application of new technologies for detection of ovarian cancer could have an important effect on public health,[1] but to achieve this goal, specific and sensitive molecular markers are essential.[1–5] This need is especially urgent in women who have a high risk of ovarian cancer due to family or personal history of cancer, and for women with a genetic predisposition to cancer due to abnormalities in predisposition genes such as *BRCA1* and *BRCA2*. There are no effective screening options for this population.

Ovarian cancer presents at a late clinical stage in more than 80% of patients,[1] and is associated with a 5-year survival of 35% in this population. By contrast, the 5-year survival for patients with stage I ovarian cancer exceeds 90%, and most patients are cured of their disease by surgery alone.[1–6] Therefore, increasing the number of women diagnosed with stage I disease should have a direct effect on the mortality and economics of this cancer without the need to change surgical or

Lancet 2002; 359: 572-577

Bias may explain 'discrimination'

Claim

- ~100% sensitivity, specificity for ovarian cancer

Problem: Compared groups: *different,* not due to cancer

- Mass spectrometry measurements done on different days in cancer specimens vs controls
- Spectrometer drifts over time; 'signal' or 'discrimination' is hardwired into results.

(Baggerly. *Bioinformatics* 2004)

Problem: Bias – Example 2

Imaging, Diagnosis, Prognosis

Diagnostic Markers for Early Detection of Ovarian Cancer

Irene Visintin,[1] Ziding Feng,[2] Gary Longton,[2] David C. Ward,[3] Ayesha B. Alvero,[1] Yinglei Lai,[4] Jeannette Tenthorey,[1] Aliza Leiser,[1] Ruben Flores-Saaib,[5] Herbert Yu,[6] Masoud Azori,[1] Thomas Rutherford,[1] Peter E. Schwartz,[1] and Gil Mor[1]

Abstract

Purpose: Early detection would significantly decrease the mortality rate of ovarian cancer. In this study, we characterize and validate the combination of six serum biomarkers that discriminate between disease-free and ovarian cancer patients with high efficiency.

Experimental Design: We analyzed 362 healthy controls and 156 newly diagnosed ovarian cancer patients. Concentrations of leptin, prolactin, osteopontin, insulin-like growth factor II, macrophage inhibitory factor, and CA-125 were determined using a multiplex, bead-based, immunoassay system. All six markers were evaluated in a training set (181 samples from the control group and 113 samples from OC patients) and a test set (181 sample control group and 43 ovarian cancer).

Results: Multiplex and ELISA exhibited the same pattern of expression for all the biomarkers. None of the biomarkers by themselves were good enough to differentiate healthy versus cancer cells. However, the combination of the six markers provided a better differentiation than CA-125. Four models with <2% classification error in training sets all had significant improvement (sensitivity 84%-98% at specificity 95%) over CA-125 (sensitivity 72% at specificity 95%) in the test set. The chosen model correctly classified 221 out of 224 specimens in the test set, with a classification accuracy of 98.7%.

Conclusions: We describe the first blood biomarker test with a sensitivity of 95.3% and a specificity of 99.4% for the detection of ovarian cancer. Six markers provided a significant improvement over CA-125 alone for ovarian cancer detection. Validation was performed with a blinded cohort. This novel multiplex platform has the potential for efficient screening in patients who are at high risk for ovarian cancer.

Clin Cancer Res 2008;14:1065

Bias may explain 'discrimination'

Claim

- ~100% sensitivity, specificity for ovarian cancer

Problem: Compared groups: *different,* not due to cancer

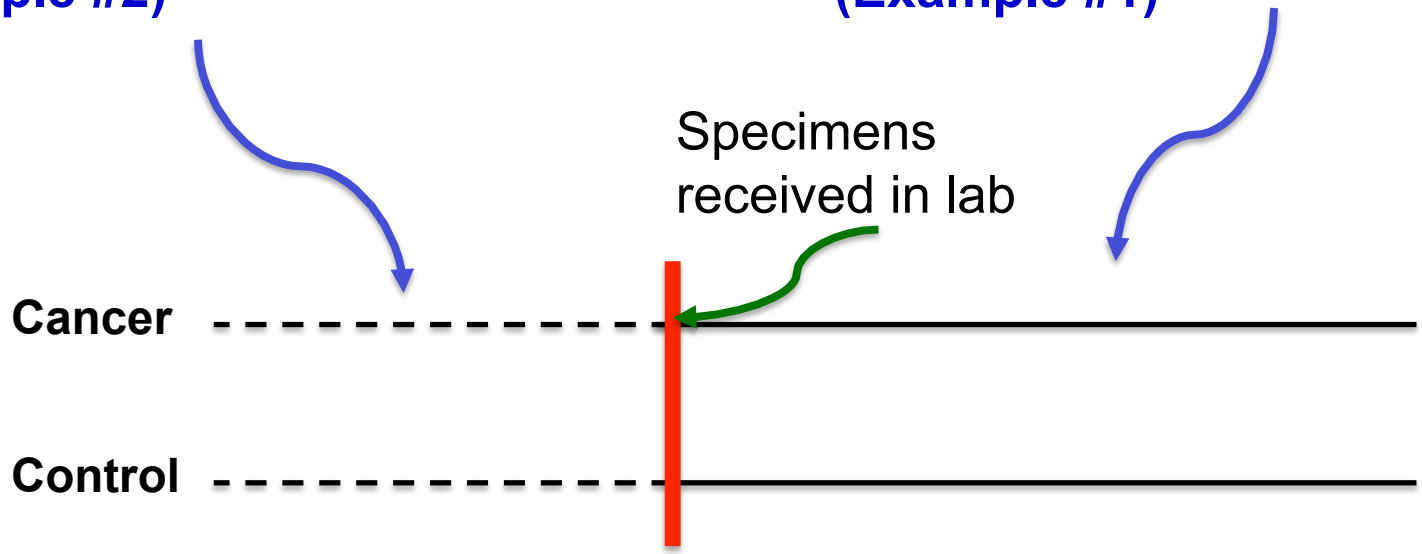
- Cancers from 'high-risk clinic' (pelvic mass)
- Controls from screening clinic
- "Stress" protein markers may differ in compared groups; bias may explain results; interpretation should be moderated.

(McIntosh. *CCR*, 2008;14:7574)

Bias may occur in different 'locations' in observational study design

Before specimens are received in lab, differences occur in demographics, collection methods, etc.
(Example #2)

After specimens are received in lab, differences occur in handling: time, place, etc.
(Example #1)



Experimental design and biospecimens

Problem

- In biomarker research, ***rate-limiting step*** is faulty study design, when bias (systematic difference between compared groups) makes results wrong and misleading

Approach

- Understand specimens are *product of* ***a study***. *Specimen collection* must be *designed* to avoid bias.

Bias as a threat to validity

Nat Rev Cancer 2005; 5:142-9

OPINION

Bias as a threat to the validity of cancer molecular-marker research

David F. Ransohoff

Abstract | Claims that molecular markers can accurately diagnose cancer have recently been disputed; some prominent results have not been reproduced and bias has been proposed to explain the original observations. As new '-omics' fields are explored to assess molecular markers for cancer, bias will increasingly be recognized as the most important 'threat to validity' that must be addressed in the design, conduct and interpretation of such research.

have been delayed by the United States Food and Drug Administration[11–13]. A number of concerns have been raised in scientific journals and the lay press[10,15–22] about whether results are reproducible and effective[10,14].

In the meantime, some observers have suggested that the pattern-recognition serum proteomics approach is not biologically plausible because some proteins or peptides might be too small to be biologically informative[16,18] or because the original results might be due to bias[19,20]. Bias can occur if the cancer and non-cancer groups

specifically, the problem of overfitting (BOX 1) — can threaten the validity of molecular-marker research[31]. This Perspectives article considers the even more important problems caused by bias.

Experimental and observational design

As summarized by Hulley and colleagues, a fundamental decision when designing studies for scientific research is "…whether to take a passive role in the events taking place in the study subjects in an observational study, or to apply an intervention and examine its effects on those events in a [randomized] clinical trial."[33] The experimental (intervention) method, provides more effective ways to deal with bias than the observational method. In clinical research, the heterogeneity of groups studied might provide particularly problematic sources of bias when groups of participants differ in ways that can affect outcome. By contrast, in a laboratory setting, the subjects might be genetically-identical cell lines

What this means for Alliance of Glycobiologists

a. As you interact with EDRN, understand where your expertise ends and others' begins (e.g. about "clinical research design").
b. You probably do not have interest/experience to design "clinical study" (obtain correct specimens) to study "cancer vs not." You probably just want to have correct specimens to apply technology/biology. *BTW the main problem is not bioinformatics or statistics.*
b. IF SO, then *utilize* EDRN experts (Karl K; Ziding Feng) to help figure out "What EDRN specimens are correct for my question."